Journal of AI and Data Mining
Vol 7, No 2, 2019, 299-309
DOI: 10.22044/JADM.2018.5321.1641

Outlier Detection for Support Vector Machine using Minimum Covariance Determinant Estimator

M. Mohammadi and M. Sarmad*

Faculty of Mathematical Sciences, Ferdowsi University of Mashhad, Mashhad, Iran.

Received 25 January 2017; Revised 14 August 2017; Accepted 12 March 2018
**Corresponding author: sarmad@um.ac.ir (M. Sarmad).*

Abstract
The purpose of this work is to identify the effective points in the performance of one of the important algorithms of data mining, namely support vector machine. The final classification decision is made based on the small portion of data called support vectors. Thus, existence of the atypical observations in the aforementioned points results in deviation from the correct decision. Therefore, the idea of Debruyne's "outlier map" is employed in this work to identify the outlying points in the SVM classification problem. However, due to the computational reasons such as convenience and rapidity, a robust Mahalanobis distance based on the minimum covariance determinant estimator is utilized. This method has a good compatibility by the data with a low-dimensional structure. In addition to the classification accuracy, the margin width is used as the criterion for the performance assessment. The larger margin is more desired due to the higher generalization ability. It should be noted that by omission of the detected outliers using the suggested outlier map, the generalization ability and accuracy of SVM are increased. This leads to the conclusion that the proposed method is very efficient in identifying the outliers. The capability of recognizing the outlying and misclassified observations for this new version of outlier map is retained similar to the older version, which is tested on the simulated and real world data.

Keywords: *Support Vector Machine, Outlying/Misclassified Points, Robust Statistics, Mahalanobis Distance, Minimum Covariance Determinant Estimator.*

1. Introduction

As a definition, an outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism [1]. Its existence is dependent upon a variety of reasons such as the variability in measurement and the experimental error. Identification of this kind of observation will lead to gain a meaningful knowledge from the data[1]-[3].

Machine learning, as an experimental science, has grown over the past 50 years with the aim of exploring whether the machine has the ability of learning or not. The variety of machine learning methods ranging from unsupervised to supervised algorithms can be widely used in various fields.

Some of the most recent works in the field of machine learning that can be pointed out are [4]-[9]. Classification is one of the machine learning tasks in which the machine learns how to separate the most similar samples from the others.

It is an important issue to identify outliers when dealing with the classification algorithms such as Support Vector Machine (SVM).

SVM is a powerful classifier due to its high generalization capability, which is used in many applications of data mining, engineering, and bioinformatics [10]. Its reputation is owing to the high classification accuracy, easiness of geometrical interpretation, and very strong fundamental theory. Another attractive property of SVM is the incorporation of the kernel function. For a non-linearly separable problem, the kernel function is used to transform the data to a higher-dimensional space.

Consider that there are n training samples $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, where $x_i \in \mathbf{R}^p$

are the input vectors and $y_i \in \{-1,+1\}$ are the class labels. The SVM objective is to find a separating hyperplane with the largest possible margin. The distance between the hyperplanes passing through the border points of both classes is known as the margin. Note that the choice of the hyperplane is crucial due to its generalization capability for classifying the future observations.
Training SVM to obtain the optimal hyperplane consists of the following minimization problem:

$$\min_{w,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \qquad (1)$$

subject to:

$$y_i\left(w^T x + b\right) \geq 1-\xi_i, \qquad \xi_i \geq 0, \quad i=1,...,n,$$

where, ξ_i is a slack variable that allows for penalized constraint violation, and C is the parameter controlling the trade-off between a large margin and a less constrained violation.
As previously mentioned, the optimal hyperplane is the one with the largest margin. Consequently, the greater certainty of correct classification is largely dependent on the margin width, which is equal to $\frac{2}{\|w\|}$. In other words, the smaller rate of misclassification of the separating hyperplane corresponds to the larger margin ([11]-[15]).
The output of a SVM optimization problem is used in the classifying function, which is given by:

$$f(x) = sign\left(w^T x + b\right) \qquad (2)$$

in which w represents the hyperplane normal vector and b is the bias term. Based on the sign of this function, the future observations can be assigned to either of the respective classes.
However, in addition to all the benefits listed above, SVM suffers from the existence of outlying points. It should be noted that in all discussions on SVM, the observations of both classes should be independently and identically distributed (i.i.d.) as an arbitrary distribution. However, this assumption does not hold in some practices. Thus the non-i.i.d. observations might become as some boundary vectors. The presence of outliers in the boundary vectors leads to shift it from the optimal place that affects the SVM performance ([13], [15]-[17]), which is illustrated in figure 1. The observations of the "First" and "Second" class are respectively depicted by the "circle" and "plus" symbols. As it can be seen in figure 1, the existence of an outlier leads to have the narrower margin.

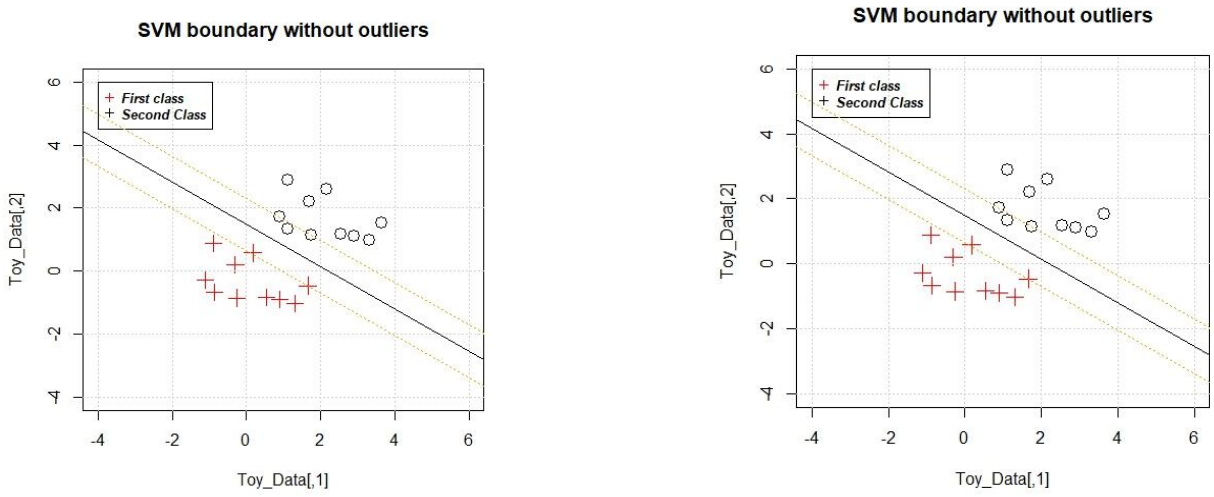

Figure 1. Sensitivity of SVM to outliers.

Thus identification of the suspected outliers should be carried out before training the SVM. Then depending on the researcher's opinion, they might be removed or a robust SVM can be used to prevent reducing the classification accuracy. The methods proposed by Zhang [15], Kou [16],

Chen [18], and Vretos [19] are some examples of robust SVM. Moreover, in the other extension of

the SVM method such as LSSVM and OCSVM, the outlier detection has been done, which can point to the recent works such as [20]-[22], [4], and [5]. Furthermore, a visual outlier detection tool, namely the "outlier map", which can be used for the classification problem, has been proposed by Debruyne [23]. It has the ability to recognize the outliers and misclassified points. The detected outlying points based on the outlier map can be removed from the dataset in order to gain a higher classification accuracy. This method can be considered as one of the eliminative methods, which deals with the outlier problem.

Note that due to the drastic effects of outliers in many practical problems, outlier detection followed by outlier removal or modification is a first step in the statistical analysis process. The detection techniques based upon the statistical approach have been divided into two separate methods, which are based either on the "projection pursuit" or the "distance".

The projection pursuit technique aims to find the most interesting projection of the data. Interesting projections are indeed the ones that deviate from the normal distribution. This method has been originally proposed by Kruskal ([24] and [25]), Switzer [26], and also Wright [27]. Actually, it can be considered as a handy tool in the multivariate data analysis to identify the outlying observations in the high-dimensional data. By high-dimensional data, we mean that the number of variables (dimension) is much larger than the number of observations $(n = p)$.

One of the most important projection pursuit methods is the Stahel-Donoho outlyingness measure, which was introduced separately by Stahel [28] and Donoho [29]. Note that the outlyingness measure can be calculated by taking the supremum of the univariate robust z-score via the projection. Simplistically, the basic idea of this method is to find the outlying points by looking at the right perspective. Since this criterion should be computed for all available directions, its computation will be highly intensive and tedious ([30] and [31]).

On the other hand, the distance-based methods such as Mahalanobis distance for the normal distributed data try to identify the outliers by comparing the distance of the particular point to the data center. The outlying points are placed on an abnormal distance from the rest of values. In the calculation of this distance, the conventional estimates of the location and scale are used. Usually the multivariate arithmetic mean and the sample covariance matrix can be used as the candidate estimators of the location and scale in this respect. It should be noted that for the multivariate normal data, this distance is distributed as χ^2_p, where p is the dimension of data. Thus the multivariate outliers are identified by the large Mahalanobis distance, which can be recognized based on the 97.5 quantile of χ^2_p as the chosen cut-off point [32]. However, due to the extreme sensitivity of the conventional estimators to the presence of outliers in the data, the robust estimators of the multivariate location vector μ and scale matrix Σ can be replaced [2].

It should be noted that in the binary classification problem, the outlier map for the high-dimensional data has been employed by Debruyne [23]; it is one of the common visual methods for the identification of outliers in the multivariate robust statistics ([2]-[34]). The Stahel-Donoho outlyingness measure has been used in the computation of the outlier map. Later on, the outlier map using Stahel-Donoho outlyingness was abbreviated as SD-SVM. As previously mentioned, it is the projection pursuit techniques for the identification of outliers in high-dimensional data. Nevertheless, sometimes the problem in hand has a low-dimensional structure, and there is no need to utilize the time-consuming methods that are used in high-dimensional data. Therefore, using a multivariate robust estimator of location and scatter, which is called the minimum covariance determinant (MCD) estimator of Rousseeuw [35], is preferred since it is computationally faster and easier than the Stahel-Donoho outlyingness measure.

In this work, based on the idea of Debruyne [23], we intend to graphically identify the outlying points using the MCD estimator. A brief explanation of this method is given in Section 3.

The MCD-based outlier map similar to the previous version [23] has the capability of discrimination of atypical points such as outlying and misclassified observations from the normal points. The difference of our proposed method with the existing technique is its fast and easiness of application due to its distance-based structure. Furthermore, a theoretical criterion for the sake of preciseness of outlier recognition rather than a vague user dependent decision-based is added. In order to assess the performance of the proposed

method, the classification accuracy and the margin width are computed, which have not been reported by Debruyne [23].

The organization of this paper is as what follows. The next section contains description of the Robust Mahalanobis distance. Section 3 is devoted to the description of SVM classifier, which is accompanied by a robust method. A detailed structure of the simulated real data, and the experimental results are given in Section 4. Section 5 presents the performance assessment of the proposed method. Finally, we conclude the paper in Section 6.

2. Robust Mahalanobis Distance (RMD)

One of the most popular outlier detection methods is the Mahalanobis distance, which takes into account the covariance of data. The Mahalanobis distance of an observation x_i with the sample mean $\bar{x}$ and sample covariance S is defined as:

$$MD(x_i) = \sqrt{(x_i - \bar{x})^T S^{-1} (x_i - \bar{x})} \quad (3)$$

In fact, it is the multivariate version of z-score, which is very sensitive to the presence of outliers [2]. However, sensitivity of the classical estimators of location and scale to the presence of outliers makes the Mahalanobis distance a non-robust measure. Plugging variants robust estimation of location and scale in the Mahalanobis distance will lead to obtain the robust version of this metric.

One of the first affine equivariant and highly robust estimators of multivariate location and scatter is the minimum covariance determinant (MCD) estimators of Rousseeuw [35]. The property of resistivity to the outlying points makes MCD a very handy tool in the robust statistics community.

Suppose that the sample consists of n independent observations, where $x_i \in \mathbf{R}^p$ for $i = 1,...,n$. The MCD's goal is to find a specific subset $H \in \{1,...,n\}$ of size h (where $\frac{n}{2} \le h \le n$) out of n, which has the smallest covariance determinant. The MCD estimator of location $\hat{\mu}_{MCD} = \frac{1}{h}\sum_{i \in H} x_i$ is the average of the aforementioned subset.

Due to the non-consistency of the MCD estimator, the consistency factor is computed. Note that calculation of this factor is not enough to render MCD an unbiased estimator, especially for a low sample size so that the swamping phenomenon will occur. Accordingly, "small sample correction factor" is calculated [36].

Therefore, the MCD scale estimator $\left(\hat{\Sigma}_{MCD}\right)$ is the covariance matrix of the subset times a factor, which is the multiplication of the consistency and finite sample correction [36]. The MCD-based robust Mahalanobis distance can be obtained by:

$$RMD(x_i) = \sqrt{(x_i - \hat{\mu}_{MCD})^T \hat{\Sigma}_{MCD}^{-1} (x_i - \hat{\mu}_{MCD})}. \quad (4)$$

The cut-off point q_η ($\eta = 0.975$ [34]) is the quantile of the chi-square distribution with p degrees of freedom. Observations associated with larger values of RMD than this cut-off point are considered to be outliers. It is worth mentioning that the MCD estimator can be computed within a reasonable time period using the FAST–MCD algorithm of Rousseeuw [37].

3. MCD-based SVM classifier

In this section, the binary SVM classification problem accompanied by a graphical outlier detection method, namely outlier map, is discussed. It is a known fact that the depiction of the data for a dimension higher than three is impossible, and there is no exception to this rule for the classification problem as well. Therefore, the graphical tool to overcome this limitation will be desirable. Firstly, for the classification problem, this plot has been proposed by Debruyne [23].

It is worthwhile to mention that the results of the outlier map exactly coincides with the SVM boundary line such that the recognized outliers and misclassified points are exactly similar in both plots. For instance, in the left panel of figure 2, the recognized outliers are distinct with their large distance to the mass of data, which is also visible in the right panel of this figure. This is also true for the misclassified points because the number of misclassified points are equal in both plots. Our proposed method has all the benefits mentioned for the outlier map. However, what distinguishes it from the outlier map is the addition of a horizontal line based on the respective cut-off point $\left(q_\eta\right)$. In contrast to Debruyne [23], our proposed method, due to the existence of a precise cut-off point, has the

advantage of identifying the real outliers. It prevents a distance-to-center decision-making that largely depends on the visual recognition.

In order to sketch the MCD-based outlier map, the amount of classifying function, the measure of outlyingness, and the cut-off point are required. As mentioned in the previous sections, in our proposed method, a criteria for the outlier recognition is the MCD-based robust Mahalanobis distance (RMD), which is computed for all observations of both classes in the first step. In the next step, instead of omission of the non-precise portion of data (50%), as proposed by Debruyne [23], the actual outlying part of data will be omitted such that for each class, RMD is calculated and the ones with the a RMD higher than the cut-off point $\left(q_\eta\right)$ will be deleted. Thus the effect of outlying points on SVM has been reduced. Then this clean part is utilized for classification by conducting the classical SVM. Also note that the computation of robust SVM is feasible in the kernel space according to Debruyne [23]. The so-called SVM based on MCD is abbreviated as MCD–SVM.

Let $\left(x_1, y_1\right), \left(x_2, y_2\right), ..., \left(x_n, y_n\right)$ be the training data for the binary classification problem. The numbers of observations in the first and second classes are, respectively, presented by n_I and n_{II}. The algorithm outline for MCD–SVM is described as follows:

- **Outlier pruning:** Compute RMD for all the observations of each class, either the first or the second one. Retain the observations with lower amounts of outlyingness (based on RMD) than q_η .

- **Training:** After the outlier pruning step, the classical SVM has been used to classify the remaining part of the data, which is illustrated by $T = T_1 \cup T_{II}$. The clean portions of the first and second classes are illustrated by T_1 and T_{II} respectively.

- **Visual representations:** Sketch the scatter plot of RMD against the amount of classifying function ((2)). The boundary line perpendicular to zero is drawn. The horizontal line passing through the cut-off point is also sketched. The samples of the two classes have been represented by different symbols, where the first class is specified by the circles and second class by the plus sign. From now on, the MCD outlier map is shortly called MCD-OM.

- **Interpretation:** The outlying and misclassified observations can be easily identified and recognized using the new version of outlier map. The 97.5 quantile of the χ^2 distribution is the criterion for outlier detection such that the observation that falls above this line can be considered as the outlier. Misclassified observations are recognized by the different colors and labels with the other observations in the respective class.

  The algorithm of this method is shown in the following.

Algorithm 1: Algorithm outline for MCD–SVM.

Input: A data matrix $x = \left[x_1, ..., x_n\right]^T$, labels vector $y = \left[-1, +1\right]_{i=1}^n$

Output: A diagram for outliers and misclassified point recognition

Initialize: $RMD = \{\ \}$, q_η *//outlyingness and* quantile of the chi-square distribution

for $i = 1, ..., n$ **do**

$RMD\left(i\right)$ = Robust Mahalanobis distance between x_i and its class center

end for

Calculate the cut-off point

Find the clean portion of data by q_η

Train a SVM model on clean train set

for $i = 1, ..., n$ **do**

Calculate $f\left(x_i\right)$

Plot $f\left(x_i\right)$ versus $RMD\left(i\right)$

end for

4. Experiment with artificial and real world data

In this section, the empirical evaluation of the proposed method is discussed through some artificial and numerical data. Due to inability to visualize the data with a dimensionality more than three $\left(p \geq 3\right)$, two-dimensional input space, i.e. $x_i \in \mathbf{R}^2$, for the artificial data is considered. It is important to note that the new version of outlier map can also be plotted for any dimension, which makes us aware of the importance of this graph.

4.1. Artificial dataset

To verify the performance of MCD–SVM, two different scenarios are considered, in which for both classes, 50 observations are generated from bivariate normal distribution with $\mu_1 = (2,2)$, $\mu_{II} = (-1,-1)$ and standard deviation equals to one.

Some simulation frameworks such as contaminated and mislabelled-contaminated have been carried out.

- **Contaminated data:** The outliers have been generated in a way that 20% of data is distributed as $MVN(\mu_I^*, I_p)$ (where $\mu_I^* = (-6,-6)$ and p is the dimension of data), which is different from the original data.
- **Mislabelled-contaminated data:** It is a mixture of contaminated and mislabelled data such that the outlying points are simultaneously mislabelled data.

The structure of simulation is given in table 1.

Table 1. Different scenarios for simulation study.

Case	Tot. no. obs.	Class I			Class II		
		Label	Mean	Total	Label	Mean	Total
Contaminated	50	*I*	μ_I	*30*	–	–	*0*
		I	μ_I^*	*20*	*II*	μ_{II}	*50*
Misclassified-Contaminated	50	*I*	μ_I	*38*	–	–	*0*
		II	μ_I^*	*12*	*II*	μ_{II}	*50*

The first part of the results obtained is devoted to the utilization of the MCD-SVM algorithm using the artificial dataset. The two-dimensional graph of the binary SVM classification is illustrated with MCD-OM in figure 1–2.

According to the left panel in figure 2, the simulated outliers are located further from the mass of dataset; they are specified by dark red color. These points are also recognized with MCD-OM. Regarding both panels in figure 2, there is no misclassified point in the data.

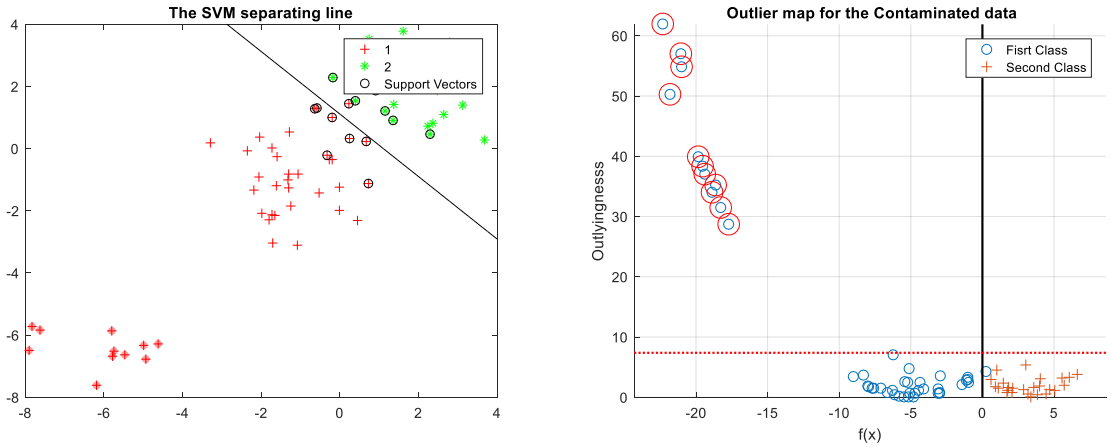

Figure 2. Plot of SVM binary classification (left) and Mahalanobis distance based on RMCD versus classifier value for contaminated (right) data. Dotted line is 97.5% quantile of chi-square distribution with p degrees of freedom.

The second scenario is depicted in figure 3. Similar to the first one, the mislabelled data that can also be considered as outliers is distinguished based on their distance, which is illustrated by dark green color. The aforementioned points are detected by the outlier map as well.

Based on the simulation scheme in table 1, the generated mislabelled data and the detected misclassified observations exactly coincide.

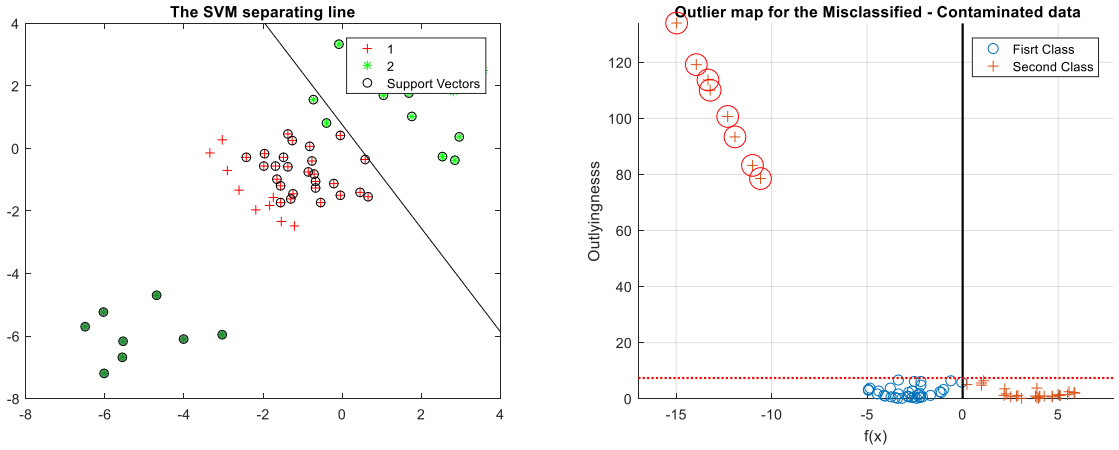

Figure 3. Plot of SVM binary classification (left) and Mahalanobis distance based on RMCD versus classifier value for Misclassified-contaminated (right) data. Dotted line is 97.5% quantile of chi-square distribution with p degrees of freedom.

From figure 3, it appears that MCD-OM using robust Mahalanobis distance has the ability to detect a more severe case, which contains the mislabelled and outlying points simultaneously.

4.2. Real dataset

The performance of the suggested techniques is assessed through some illustrative real world datasets from UCI machine learning repository[1]. The famous Ionosphere, Diabetes, Iris (for the compatibility issue with the binary classification, two classes of these datasets are considered) and Sonar data have been used.

• **Ionosphere data:** This dataset with 351 observations and 34 real-valued predictors has been collected by Sigillito et al. [38]. The label is categorical with two levels as 'g', which denotes good radar returns, and 'b', which indicates the bad radar returns. Based on the 34 attributes, the goal is to predict a good or bad return.

• **Diabetes data:** This dataset is about measuring eight different attributes of female patients of Pima Indian heritage, which are at least 21 years old. The class labels indicate whether the patients have a sight of disease or not.

• **Iris data:** The dataset actually contains 3 classes of 50 instances (Setosa, Versicolour, and Virginica), where each class refers to a type of iris flower. In this paper, we only consider two classes, whose labels are Versicolour and Virginica. The dataset contains measurements of the four variables sepal length and width and petal length and width.

• **Sonar data:** The sonar signal has been collected in this dataset, which contains 60 continuous attributes. The aim is to distinguish between signals bounced off a rock or a metal cylinder.

MCD-OM for the above-mentioned datasets using linear and Gaussian kernels is presented in figure 4-7.

The number of observations that are tagged as outlying points in the Ionosphere data is almost high, present in both panels in figure 4.

MCD-OM using linear kernel indicates a lower number of misclassified observations than the Gaussian one. Accordingly, it is preferable to use the linear kernel.

The outlying and misclassified points, which have almost allocated to themselves 30% of the data, are illustrated in both panels of figure 5. The detected outlier based on MCD-OM can be removed or eliminated.

Figure 6 presents one interesting point placed very far from the data, which is shown by its high outlyingness. The other outliers above the cut-off point are almost in the same range. In this data, there are only two misclassified observations that are very near to the boundary line. MCD-OM using Gaussian kernel shows six outliers and three misclassified outliers as well.

Figure 7 indicates that the numbers of outlying points in both panels are similar but the number of misclassified observations in MCD-OM using the Gaussian kernel is higher than the linear one.

[1] http://archive.ics.uci.edu/ml/

Consequently, the linear kernel is the desired kernel.

According to MCD-OM using the Gaussian kernel in figure 4-7, almost a large part of the data is placed around zero. It is an expected phenomenon, as the Gaussian kernel tries to classify the data as much as possible. It tries to get closer to the data so the corresponding classifying function $\left(f\left(x \right) \right)$ is close to zero.

As a remarkable feature for this chart, it should be noted that the SVM plot cannot be plotted for the $p \geq 3$ dimensions. Thus the only method capable of recognizing the atypical points is MCD-OM.

4.3. Performance assessment

In order to check the performance of the proposed method, the classification accuracy and also the margin width are computed. As mentioned earlier, the hyperplane with the largest separating margin ensures the maximization of the generalization ability of SVM. Thus the separating hyperplane with the largest margin is desired. Here, in order to test whether the detected outliers by MCD-OM are effective observations on the generalization ability of SVM or not, the SVM margin as well as the classification accuracy are computed for the simulated and real data.

Table 2 presents the accuracy of classification and margin width before and after omission of the detected outliers by SD-OM and MCD-OM. The results obtained are compared with classical SVM. It is obvious that by removing the detected outliers by MCD-OM, the margin, and subsequently, generalization ability are increased. This also happen for the classification accuracy. In fact, the experiments are repeated 10 times, and the average accuracy of the methods along with the average of margin width are presented in tables 2 and 3. It is obvious that our model has the best performance (a higher classification accuracy) compared with classic SVM and SD-OM. It also shows that our proposed method almost outperforms the other methods in terms of margin width, with a lower standard division. It should be noted that the above-mentioned measures have also been computed for the real world dataset using the Gaussian kernel. The results obtained are depicted in table 3.

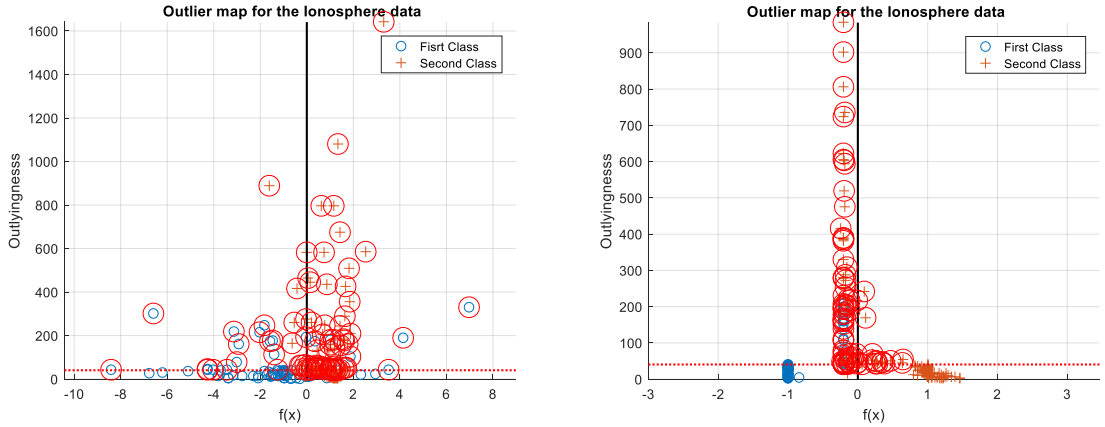

Figure 4. Plot of Mahalanobis distance based on MCD versus classifier value for Ionosphere dataset using linear (left panel) and Gaussian (right panel) kernels.

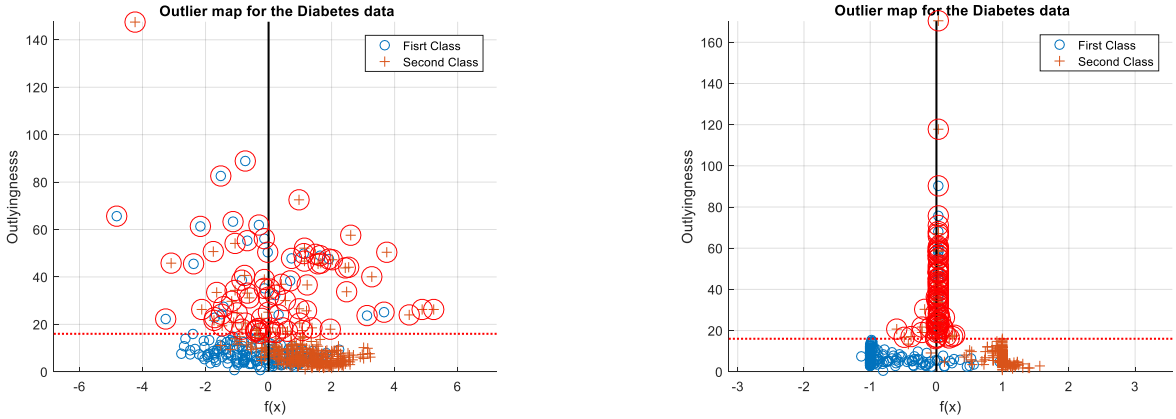

Figure 5. Plot of Mahalanobis distance based on MCD versus classifier value for Diabetes dataset using linear (left panel) and Gaussian (right panel) kernels.

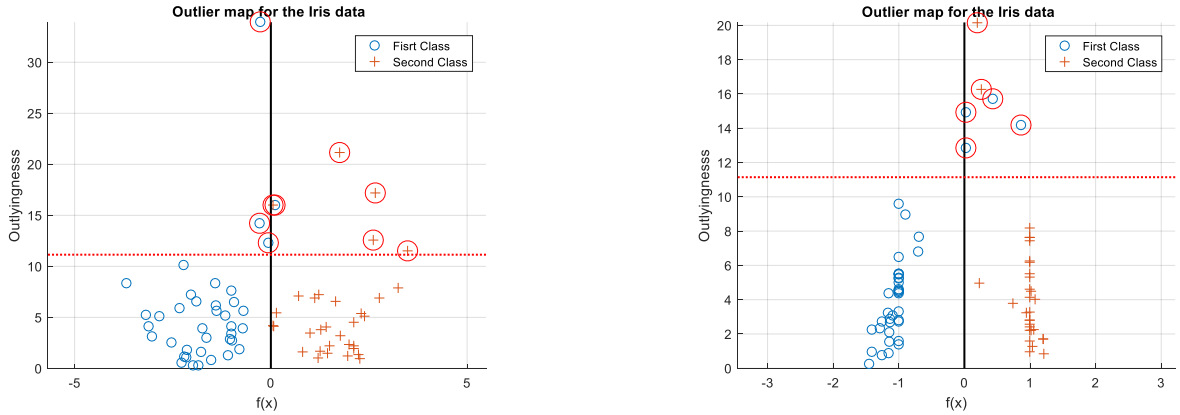

Figure 6. Plot of Mahalanobis distance based on MCD versus classifier value for two classes Iris dataset using linear (left panel) and Gaussian (right panel) kernels.

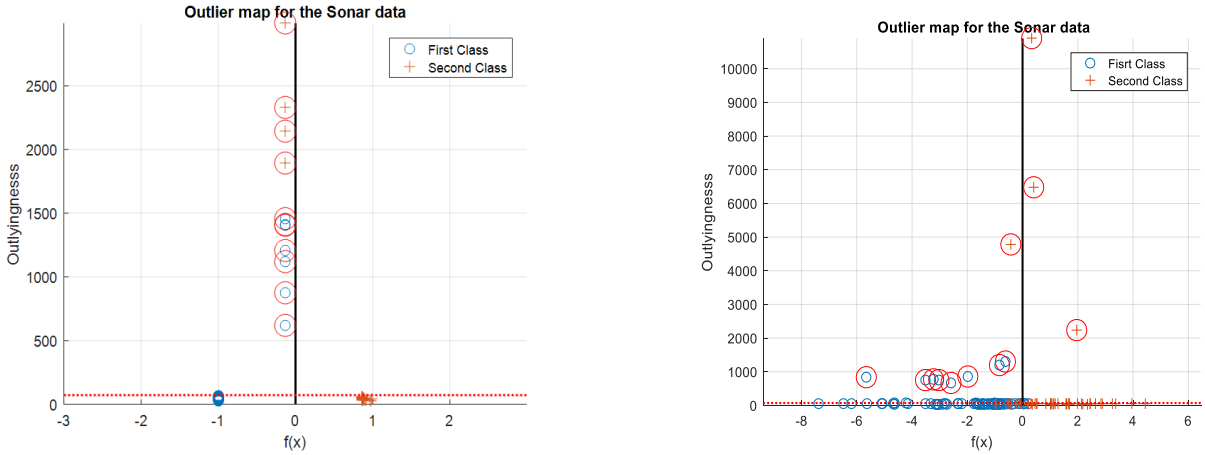

Figure 7. Plot of Mahalanobis distance based on MCD versus classifier value for two classes Sonar dataset using linear (left panel) and Gaussian (right panel) kernels

Table 2. Classification accuracy and margin width before and after outlier removal using linear kernel.

Dataset	Classification accuracy			Margin width		
	SVM	SD	MCD	SVM	SD	MCD
Contaminated	97.00±(0.02)	94.33±(0.05)	**98.33±(0.02)**	0.61±(0.01)	**0.89±(0.18)**	0.63±(0.01)
Misclassified-Contaminated	76.33±(0.06)	81.33±(90.08)	**82.66±(0.05)**	1.15±(0.30)	1.21±(0.23)	0.37±(034)
Ionosphere	87.64±(0.03)	81.79±(0.04)	**87.93±(0.02)**	0.22±(0.01)	0.23±(0.01)	**0.47±(0.01)**
Diabetes	75.45±(0.01)	72.07±(0.05)	**75.93±(0.01)**	0.24±(0.01)	0.28±(0.01)	**0.37±(0.20)**
Iris	97.01±(0.03)	88.33±(0.06)	**97.66±(0.02)**	0.31±(0.20)	0.34±(0.01)	**0.73+(0.16)**
Sonar	76.01±(0.03)	66.50±(0.06)	**77.78±(0.03)**	0.19±(0.01)	**0.71±(0.29)**	0.22±(0.01)

Table 3. Classification accuracy and margin width before and after outlier removal using Gaussian kernel.

Dataset	Classification accuracy			Margin width		
	SVM	SD	MCD	SVM	SD	MCD
Ionosphere	72.35±(0.12)	63.77±(0.03)	**81.13±(0.02)**	0.07±(0.01)	**0.23±(0.01)**	0.08±(0.01)
Diabetes	66.19±(0.02)	65.36±(0.02)	66.19±(0.02)	0.24±(0.01)	0.28±(0.02)	**0.37±(0.01)**
Iris	93.00±(0.05)	63.00±(0.10)	**93.67±(0.04)**	0.20±(0.08)	0.22±(0.02)	**0.51+(0.01)**
Sonar	51.58±(0.06)	52.06±(0.05)	**51.73±(0.01)**	0.08±(0.01)	**0.42±(0.15)**	0.09±(0.01)

5. Conclusion

In this paper, a statistical preprocessing approach for outlier pruning was introduced, which is a critical step in the data mining process. For the purpose of remediation of the problem caused by the presence of outliers on SVM, the robust methods were applied. In this work, the binary SVM classification problem was accompanied by the visual outlier detection method, namely the outlier map. High breakdown robust methods, namely Mahalanobis distance based on the minimum covariance determinant, were utilized in the computation of the outlier map. The performance of the proposed robust SVM algorithm was assessed by some artificial and real datasets. As an interesting result of this work worth to mention, omission of the detected outliers by MCD-OM resulted in increase in the SVM margin width and consequently, increase in the generalization ability of it simultaneously. Classification accuracy, as the other measure of preciseness of classification was calculated and the obtained results were reported in tables 2 and 3. The results confirmed the superiority of MCD-SVM over SD-SVM and classical SVM.

References

[1] Hawkins, D. M. (1980). Identification of outliers, London, New York, Chapman and Hall.

[2] Rousseeuw, P. J., & Van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. Journal of the American Statistical association, vol. 85, no. 411, pp. 633-639.

[3] Ahmadi Livani, M., Alikhany, M., & Yadollahzadeh Tabari, M. (2013). Outlier detection in wireless sensor networks using distributed principal component analysis. Journal of AI and Data Mining, vol. 1, no. 1, pp. 1-11.

[4] Xiao, Y., Wang, H., & Xu, W. (2017). Ramp Loss based robust one-class SVM. Pattern Recognition Letters, vol. 85, pp. 15-20.

[5] Xu, G., Cao, Z., Hu, B.-G., & Principe, J. C. (2017). Robust support vector machines based on the rescaled hinge loss function. Pattern Recognition, vol. 63, pp. 139-148.

[6] Jiao, L., Shang, F., Wang, F., & Liu, Y. (2012). Fast semi-supervised clustering with enhanced spectral embedding. Pattern Recognition, vol. 45, no. 12, pp. 4358-4369.

[7] Shang, R., Zhang, Z., Jiao, L., Wang, W., & Yang, S. (2016). Global discriminative-based nonnegative spectral clustering. Pattern Recognition, vol. 55, pp. 172-182.

[8] Shang, R., Wang, W., Stolkin, R., & Jiao, L. (2016). Subspace learning-based graph regularized feature selection. Knowledge-Based Systems, vol. 112, 152-165.

[9] Shang, R., Wang, W., Stolkin, R., & Jiao, L. (2018). Non-negative spectral learning and sparse regression-based dual-graph regularized feature selection. IEEE transactions on cybernetics, vol. 48, no. 2, pp. 793-806.

[10] Lee, Y. (2010). Support vector machines for classification: a statistical portrait. In Statistical Methods in Molecular Biology, Springer, pp. 347-368.

[11] Vapnik, V., & Kotz, S. (2006). Estimation of dependences based on empirical data: empirical inference science (information science and statistics), Springer-Verlag Berlin, Heidelberg.

[12] Vapnik, V. (2013). The nature of statistical learning theory: Springer science & business media, Springer-Verlag New York.

[13] Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. Paper presented at the Proceedings of the fifth annual workshop on Computational learning theory, Pittsburgh, Pennsylvania, USA, 1992.

[14] Shawe-Taylor, J. (1998). Classification accuracy based on observed margin. Algorithmica, vol. 22, no. 1-2, pp. 157-172.

[15] Zhang, X. (1999). Using class-center vectors to build support vector machines. Paper presented at the Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop.

[16] Kou, Z., Xu, J., Zhang, X., & Ji, L. (2001). An improved support vector machine using class-median vectors. Paper presented at the Proc of 8th Intl Conf on Neural Information Processing, Shanghai, China, 2001.

[17] Song, Q., Hu, W., & Xie, W. (2002). Robust support vector machine with bullet hole image classification. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 32, no. 4, pp. 440-448.

[18] Chen, J. H. (2004). M-estimator based robust kernels for support vector machines. In Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on (Vol. 1, pp. 168-171). IEEE, Washington, DC, USA, 2004.

[19] Vretos, N., Tefas, A., & Pitas, I. (2013). Using robust dispersion estimation in support vector machines. Pattern Recognition, vol. 46, no. 12, pp. 3441-3451.

[20] Feng, Y., Yang, Y., Huang, X., Mehrkanoon, S., & Suykens, J. A. (2016). Robust support vector machines for classification with nonconvex and smooth losses. Neural computation, vol. 28, no. 6, pp. 1217-1247.

[21] Prayoonpitak, T., & Wongsa, S. (2017). A Robust One-Class Support Vector Machine Using Gaussian-Based Penalty Factor and Its Application to Fault Detection. International Journal of Materials, Mechanics and Manufacturing vol. 5, no. 3, pp. 146-152.

[22] Chen, L., & Zhou, S. (2018). Sparse algorithm for robust LSSVM in primal space. Neurocomputing, 275, 2880-2891.

[23] Debruyne, M. (2009). An outlier map for support vector machine classification. The Annals of Applied Statistics, vol. 3, no. 4, pp. 1566-1580.

[24] Kruskal, J. B. (1969). Toward a ractical method which helps uncover the tructure of a set of multivariate observations by finding the linear transformation which optimizes a new "Index of condensation". Paper presented at the Statistical Computation, Statistical Computation, New York : Academic Press, pp.427-440.

[25] Kruskal, J. B. (1972). Linear transformation of multivariate data to reveal clustering. Multidimensional scaling: theory and applications in the behavioral sciences, vol. 1, pp. 181-191.

[26] Switzer P. (1970) Numerical Classification. In: Merriam D.F. (eds) Geostatistics. Computer Applications in the Earth Sciences. Springer, Boston, MA, pp. 31-43.

[27] Wright, R. M., & Switzer, P. (1971). Numerical classification applied to certain Jamaican Eocene nummulitids. Journal of the International Association for Mathematical Geology, vol. 3, no. 3, pp. 297-311.

[28] Stahel, W. A. (1981). Robuste schätzungen: infinitesimale optimalität und schätzungen von kovarianzmatrizen. ETH Zurich.

[29] Donoho, D. L. (1982). Breakdown properties of multivariate location estimators. Technical report, Harvard University, Boston. URL http://www-stat. stanford. edu/~ donoho/Reports/Oldies/BPMLE. pdf.

[30] Chenouri, S. E., Steiner, S. H., & Variyath, A. M. (2009). A multivariate robust control chart for individual observations. Journal of Quality Technology, vol. 41, no. 3, pp. 259-271.

[31] Lawrence, D. E., Birch, J. B., & Chen, Y. (2014). Cluster-Based Bounded Influence Regression. Quality and Reliability Engineering International, vol. 30, no. 1, pp. 97-109.

[32] Johnson, R. A., & Wichern, D. W. (2004). Multivariate analysis. Encyclopedia of Statistical Sciences, 8.

[33] Hubert, M., & Engelen, S. (2004). Robust PCA and classification in biosciences. Bioinformatics, vol. 20, no. 11, pp. 1728-1736.

[34] Hubert, M., Rousseeuw, P. J., & Vanden Branden, K. (2005). ROBPCA: a new approach to robust principal component analysis. Technometrics, vol. 47, no. 1, pp. 64-79.

[35] Rousseeuw, P. J. (1984). Least median of squares regression. Journal of the American Statistical association, vol. 79, no. 388, pp. 871-880.

[36] Pison, G., Van Aelst, S., & Willems, G. (2003). Small sample corrections for LTS and MCD. METRIKA, vol. 55, no. 1-2, pp.111–123.

[37] Rousseeuw, P. J., & Driessen, K. V. (1999). A fast algorithm for the minimum covariance determinant estimator. Technometrics, vol. 41, no. 3, pp. 212-223.

[38] Sigillito, V. G., Wing, S. P., Hutton, L. V., & Baker, K. B. (1989). Classification of radar returns from the ionosphere using neural networks. Johns Hopkins APL Technical Digest, vol. 10, no. 3, pp. 262-266.

نشریه هوش مصنوعی و داده کاوی

شناسایی مشاهدات دورافتاده موثر بر ماشین بردار پشتیبان با استفاده از برآوردگر مینیمم دترمینان کوواریانس

ماندانا محمدی و مجید سرمد*

دانشکده علوم ریاضی، دانشگاه فردوسی مشهد، مشهد، ایران.

ارسال ۲۰۱۷/۰۱/۲۵؛ ارسال ۲۰۱۷/۰۸/۱۴؛ پذیرش ۲۰۱۸/۰۳/۱۲

چکیده:

هدف این مقاله شناسایی نقاط تاثیر گذار بر عملکرد یکی از مهم‌ترین الگوریتم‌های داده‌کاوی به‌نام ماشین بردار پشتیبان است. نتیجه نهایی طبقه‌بندی بر اساس مجموعه محدودی از نقاط به‌نام بردارهای پشتیبان می‌باشد که وجود مشاهدات غیر عادی در آن‌ها باعث ایجاد تغییر اساسی در نتیجه طبقه‌بندی می‌شود. بنابراین از ایده نمودار شناسایی مشاهدات دورافتاده‌ی دبروین برای شناسایی این‌گونه مشاهدات استفاده شده است. در این مقاله، بنابه دلایل محاسباتی مانند سهولت و سرعت بالا از یک فاصله ماهالانوبیس استوار بر مبنای برآوردگر مینیمم دترمینان کوواریانس استفاده شده است. شایان ذکر است که این روش برای داده‌هایی که از لحاظ ساختاری دارای ابعاد کمی هستند، مناسب است. برای بررسی عملکرد ماشین بردار پشتیبان، از پهنای باند به‌عنوان معیار دیگری علاوه برای دقت طبقه‌بندی استفاده شده است که پهن‌تر بودن آن به‌دلیل ارتباط مستقیم با بالاتر بودن خاصیت تعمیم‌پذیری، دارای مطلوبیت بیشتری است. گفتنی است که با کنار گذاشتن مشاهدات شناسایی شده توسط این نمودار، معیارهای پهنای باند و دقت طبقه‌بندی افزایش می‌یابد که نشان‌دهنده توانایی تشخیص درست مشاهداتی است که باعث کاهش عملکرد ماشین بردار پشتیبان می‌شوند. توانایی تشخیص مشاهدات غیر عادی، اعم از مشاهدات دورافتاده و مشاهدات به اشتباه طبقه‌بندی شده نسخه جدید نمودار مانند نسخه قدیمی است که با استفاده از داده‌های واقعی و شبیه سازی شده مورد آزمایش و تایید قرار گرفت.

کلمات کلیدی: ماشین بردار پشتیبان، مشاهدات دورافتاده/ به اشتباه طبقه‌بندی شده، آمار استوار فاصله ماهالانوبیس، برآوردگر مینیمم دترمینان کوواریانس.